DOCUMENT RESUME

ED 435 701 TM 030 332

AUTHOR Fan, Xitao; Wang, Lin
TITLE Practical Guidelines for Evaluating Sampling Designs in Survey Studies.
PUB DATE 1999-11-00
NOTE 25p.; Paper presented at the Annual Meeting of the American Evaluation Association (14th, Orlando, FL, November 3-6, 1999).
PUB TYPE Guides - Non-Classroom (055) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Evaluation Methods; *Research Design; Research Methodology; *Sampling; *Surveys

ABSTRACT

The popularity of sample surveys in evaluation and research makes it necessary for consumers to tell a good survey from a poor one. Several sources were identified that gave advice on how to evaluate a sample design used in a survey study. The sources are either too limited or too extensive to be useful practically. The purpose of this paper is to recommend six important yet practical guidelines for evaluating the quality of a sample design in a survey study. The six guidelines are: (1) a clearly specified population; (2) an explicitly stated unit of analysis; (3) a specification of determining a desired sample size; (4) an informative description of the sample selection procedures; (5) a description of response rate and nonresponse treatment; and (6) demonstration of appropriate estimation procedures. For each guideline, discussion is focused on definitions, problems found in the literature, and the consequences of the problems. (Contains 30 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Evaluate Sample Design.wpd

ED 435 701

Practical Guidelines for Evaluating Sampling Designs in Survey Studies

Xitao Fan

Utah State University

Lin Wang

American College Testing

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Xitao Fan

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Running Head: Evaluating Survey Sample Design

Send correspondence about this paper to:

Xitao Fan, Ph.D.
Department of Psychology
Utah State University
Logan, Utah 84322-2810

Phone: (435)797-1451
Fax: (435)797-1448
E-mail: fafan@cc.usu.edu
Web: http://cc.usu.edu/~fafan/

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM030332

Paper presented at 1999 American Evaluation Association Conference, November 3-6, Orlando, Florida (Session No.: 767).

BEST COPY AVAILABLE

ABSTRACT

The popularity of sample survey in evaluation and research makes it necessary for consumers to tell a good study from a poor one. Several sources were identified that gave advice on how to evaluate a sample design used in a survey study. The sources are either limited or too extensive to be practically useful. The purpose of this paper is to recommend six important yet practical guidelines in evaluating the quality of a sample design in a survey study. The six guidelines are 1) a clearly specified population, 2) an explicitly stated unit of analysis, 3) a specification of determining a desired sample size, 4) an informative description of the sample selection procedures, 5) a description of response rate and non-response treatment, and 6) demonstration of appropriate estimation procedures. For each guideline, discussion is focused on definitions, problems that are found in literature, and consequences of the problems.

Survey research is widely used and reported in evaluation and research studies. Since a survey is often not administered to everybody in a group of interest (population), a survey is actually a sample survey, i.e., only a part of the intended population is usually involved. Researchers using the sample survey method hope to gain knowledge about the population based on the results from the sample in hand. When such results are reported, consumers (e.g., policy makers, practitioners, other evaluators and researchers) expect that the results from the sample can be generalized to the population of interest.

Because sample surveys work on partial information from a population, requirements have been established for conducting a sample survey to ensure the quality of information obtained and the validity of generalizing this information to the population. Of those various requirements, the ones concerning sample design deserve great attention. Typically, these sample design requirements involve sample selection and estimation procedures (Kish, 1965), which provides technical assurance to the validity and generalizability of the survey findings for the intended use.

Just as in many other aspects of evaluation and research, sample design in survey research often has a variety of flaws in their sample design part (Miskel & Sandlin, 1981; Pena & Henderson, 1986; Permut, Michel, & Joseph, 1976; Wang, 1996; Wang & McNamara, 1997). Whatever the findings or conclusions drawn from those studies are prone to problems with regard to the validity and generalizability issue. It is, therefore, important that the consumers of these studies have adequate knowledge and skills to evaluate the usefulness of the information so as not to fall victim of blind belief in any misinformation. Recommendations about how to design a good sample survey are provided in many texts; these recommendations can also serve as general guidelines for evaluating survey research. Criteria specifically used in evaluating sample survey are, however, found in only a few sources, and

BEST COPY AVAILABLE

these criteria range from broad categories (Miskel & Sandlin, 1981; Permut et al, 1976) to detailed checklists (Jaeger, 1988; McNamara, 1994).

In the evaluation work of both Miskel and Sandlin (1981) and Permut et al. (1976), the broad criteria included population specification, unit of analysis, sampling frame, response rate, and selection procedures. These are very important criteria for sample design evaluation, especially those concerning population, unit of analysis and response rate. However, these criteria appear to be confined to the selection part of sample design only, whereas the estimation part, as defined by Kish (1965), is often not mentioned at all. Estimation is closely related to selection in a sample design. For example, both unit of analysis and response rate will have bearing on what analysis to be used in estimation, and population specification will determine how the results will be interpreted. It is, therefore, necessary to include estimation procedures in the evaluation criteria.

The checklists from Jaeger (1988) and McNamara (1994) are both extensive, with 26 and 100 items, respectively. The items cover a wide range of aspects of survey research, from research question and instrumentation to survey finding interpretation and recommendations. These checklists are valuable sources of information to researchers and people learning to conduct survey research. This is particularly true of McNamara's list, which is so detailed that following the list will certainly strengthen the sample design. The downside of these checklists, however, is that the lists are too long and may become overwhelming for practitioners. Although the items in the checklists are important, some are more important than others with regard to their impact on the outcome of a sample survey study. Accordingly, from the perspective of a consumer, one needs to know the most important or crucial criteria that can be conveniently applied to practical use in evaluating survey research literature.

It is helpful for survey research consumers to know what problems one may expect to find in a sample design and what the possible consequences of such problems are. For example, one should be

able to tell from the description of a sample design whether the reported sample size may be problematic, and if it is, what impact this problem may have on the findings. Equipped with such knowledge, one may then make informed judgment about how to use the information from this survey research in an appropriate manner. Such discussion, however, is not directly available in the texts presenting the checklists for survey evaluation.

The purpose of this paper is to provide a focused guide that highlights what we deem as key criteria for evaluating sample design in evaluation and research applications. Here we consider six guidelines that can be conveniently used by professionals (e.g., evaluators, researchers, policy makers, administrators) to evaluate the quality of sample designs. These are:

1. a clearly specified target population and survey population,
2. an explicit statement of unit of analysis in light of the research questions,
3. a specification of a desired sample size and how this is determined,
4. an informative description of the sample selection procedures,
5. a description of response rate with information about non-response situation, and
6. demonstration of appropriate estimation and data analysis according to the selection strategies used, including the treatment of possible non-response bias problem and generalization of findings.

In presenting each guideline, three aspects will be discussed. First, a criterion will be described or defined within the context of sample survey. Second, what problems might be encountered that are related to this criterion, and where appropriate, examples drawn from an evaluation study by Wang (1996) will be used for illustration. Finally, possible consequences of not meeting the criterion will be explored.

BEST COPY AVAILABLE

The Six Guidelines

Guideline One: Population Specification

The intended population is important in survey research because it is the characteristics of the population that we are interested in, and are trying to figure out by conducting a sample survey. In survey research, a population is usually a finite population that "consists of a known, finite number N of units--such as people or plots of ground. With each unit is associated a value of a variable of interest" (Thompson, 1992, p. 2). A distinction is often made between a target population and a survey population because the two may not be the same in reality (Gall, Borg, & Gall, 1996; Kish, 1965; Sudman, 1976).

A target population (also called "the inference space" in statistical theory) is the ideal population of interest, or the desired population, or "the total finite population about which we require information" (Barnett, 1991, p. 8). A target population must be defined in terms of content, units, extent, and time (Kish, 1965). By units it means whether a target population is a population of individuals, schools, households, etc. (Sudman, 1976). These units are also defined as sampling units, which can be individuals or aggregates such as schools or classes. For example, if we want to estimate the proportion of public school principals in the United States who hold graduate degrees, the target population then should consist of every single public school principal in the country. The sampling units are the individual principals. In practice, however, the target population is often difficult to access due to various practical constraints, and consequently, it has to be modified and reduced to a survey population. The importance of having a target population is that we know what has been excluded to obtain the survey population, and we can therefore assess the consequences of such exclusion while interpreting the findings (Kalton, 1983b).

A survey population (also called study population, experimentally accessible population, operational population, or sampled population) is the actual population studied. In the example above, the survey population may not include principals of schools located in some remote or thinly populated areas so as to reduce the cost of administering the survey. Namely, "resources and feasibility determine that some members of a target population have to be excluded " (Fink, 1995). The remaining members of the target population then constitute a survey population. The difference between the two types of population requires additional consideration in making inference and generalization. Since a sample is taken from a survey population, the statistical inference can be made only to the survey population. The generalization of results from the survey population to the target population is not based on any statistical decision but, instead, on substantive knowledge and subjective judgment (Babbie, 1986; Cochran, 1977; Gall, Borg, & Gall, 1996; Jolliffe, 1986; Sudman, 1976). In other words, population validity has to be established to justify appropriate generalizations from a survey population to a target population. Any relevant differences between the two populations need to be discussed or accounted for. A simple way to make this point might be to use Jaeger's (1988) comment that a target population reflects a researcher's desires and a survey population (or a sampling frame) defines reality.

Depending on the actual scope and nature of a survey study, the two types of population defined above can be the same, because it is sometimes feasible to sample directly from a target population without any exclusion. This can happen when a researcher conducts survey in a special population that can be controlled with some ease. For instance, when an evaluator has to use a sample survey to obtain certain information from a population of full time school teachers in a district, the target population includes all the full time teachers as defined by that district. This target population is typically limited in size and easily accessible; the sample can be taken from a sampling frame built from the roster of all the teachers in the district. Therefore, no difference exists between the target

BEST COPY AVAILABLE

population and survey population. In many other situations, however, compromises or modifications have to be made between a survey population and a target population.

It has been observed that many survey-based studies in research literature fail to report or adequately describe what populations are used. For example, in an evaluation of the survey studies published in a peer-reviewed journal in 1970s, Miskel and Sandlin (1981) reported that only seven of the 23 studies evaluated (30.4%) specified their populations. Wang (1996) examined the sample designs in the survey studies published by the same journal during 1980s and 1990s and found that only 11 of the 53 sample designs evaluated (20.8%) mentioned their populations.

Where populations are not clearly specified, it is left to the readers to figure out what populations may have been involved. The missing information about the population has at least two consequences. One is that it creates a problem for readers to have a clear idea about what population was actually studied, and as a result, it becomes difficult for readers to evaluate whether a sample was good enough to be the fair representation of the population. The other consequence is that the scope of generalization of any findings becomes a questionable issue, since no particular population has been specified. Although authors of such studies often discuss the implication or significance of their findings, which often involves generalization, it is difficult to justify such a discussion in the absence of a clearly defined population. This problem will certainly affect the validity of a study no matter how well the other parts of the study are.

Guideline Two: Unit of Analysis Specification

The specification of unit of analysis in a sample design is necessary because this specification dictates outcomes of data analysis. Descriptions of unit of analysis can be found in many texts (Babbie, 1986; Gall, Borg, & Gall, 1996; Fink, 1995). Babbie (1986) defined units of analysis as "those units

that we initially describe for the ultimate purpose of aggregating their characteristics in order to describe some larger group or explain some abstract phenomenon" (p. 74). The importance of specifying units of analysis has been strongly emphasized. "When the unit of analysis is not so clear, however, it is absolutely essential to determine what it is; otherwise, you will be unable to determine what observations are to be made about whom or what" (Babbie, 1986, p. 76). This statement applies to many disciplines in social and behavioral sciences which share some common research methods.

Specifying the unit of analysis is done in the design stage based on whether the purposes of a study is to investigate some phenomena at the individual or the group level. Gall, Borg and Gall (1996) has made it clear that, in education, activities occur at different levels: individual students, groups of special interest, classroom, school, etc.. Researchers need to decide which of these levels contains the phenomena of interest to them, and to develop an explanation for why they are interested in these phenomena.

For example, if a researcher has a good reason to study teacher's job satisfaction, individual teachers should be the unit of analysis even though schools may have to be selected first before teachers can be sampled from the schools. The schools are not the units of analysis because they do not constitute the interest of the study. In another example, if an evaluator plans to investigate what factors are related to school effectiveness, the unit of analysis is school, not individuals like principals, teachers, or students, even though raw data must be collected from individuals working at a school. The individual data is usually aggregated to the level (school) as defined by the unit of analysis before further analysis is conducted.

Unit of analysis is very often overlooked. Again, in Miskel and Sandlin's (1981) report, they only found 7 out of 23 studies that contained information about units of analysis. In Wang's (1996)

BEST COPY AVAILABLE

evaluation, only 23 out of 53 sample designs had such information. In other words, it appears that at least more than half of the studies did not say what units of analysis had been considered in their designs and analysis.

Failure to specify the unit of analysis may not lead to serious problems in a simple study that only involves data from one level; the correspondence between research questions and unit of analysis is sometimes apparent even if nothing is said about the latter. In a more extended study that involves sampling at several levels, however, problems may arise if unit of analysis is not given due attention. A typical problem a researcher may have is to do the analysis on a wrong level.

For example, when we study a sample of 20 classes, with five students sampled from each class of 30, to investigate students' expectation of their teachers, class is the right unit of analysis and the data from five students should be aggregated to yield a data point for their class. In the subsequent analysis, the sample size is 20, the number of classes used. If the analysis is conducted using individual students as unit of analysis, the sample size would be 100. The findings would be distorted and not appropriate for the interest of research. Because of the exaggerated sample size and the inflated statistical power, trivial differences among the classes may now become statistically significant. Conclusions based on the incorrect analysis will fall victim to the misuse of the unit of analysis in the study.

Traditionally, there is usually only one level of unit of analysis, because one level of unit of analysis is what conventional analytic strategies (e.g., ANOVA) can handle. However, recent developments in some new quantitative analysis strategies (e.g., hierarchical linear modeling, or HLM) (Bryk & Raudenbush, 1992), and the implementation of such strategies in statistical software, have made it possible that multiple levels of units of analysis can be specified and analyzed.

Guideline Three: Sample Size Determination

BEST COPY AVAILABLE

Another important task in sample design is determining an appropriate sample size, and this information should be reported. Sample size decision is primarily based on desired analysis and efficiency. The analysis refers to parameter estimation and hypothesis testing, the two fundamental components of statistical analysis. Sample size plays an important role in both cases in that sample size influences the margin of error, or degree of precision in estimation, and affects statistical power in hypothesis testing (Cohen, 1988; Fink, 1995; McNamara, 1994; Sheaffer, Mendenhall, & Ott, 1990). Efficiency is about obtaining the same amount of information with as small a sample as possible. As Jaeger (1988) put it, "Choosing the best possible sampling method is one important way of increasing the efficiency of a survey, thus reducing costs without sacrificing quality or precision" (p. 317).

To illustrate this point, Jaeger (1988) gave an example that, to estimate the average score of 1200 students in a school system, the most efficient procedure would require testing only 25 students, whereas the least efficient sampling and estimation procedures would have to test 1,041 students. Methods for calculating desired sample sizes can be found in many statistics and research method books, among which an introductory sampling text by Sheaffer et al. (1990), a statistical power handbook by Cohen (1988), and a text by McNamara (1994) appear to be fairly comprehensive and easy to use.

Specifying sample size decisions appears to be another weak area in that reported survey studies in the literature tend to give only the achieved sample size used in analysis, but provide little information about how a planned sample size is determined. For example, in the 53 sample designs evaluated by Wang (1996), only two mentioned their sample size decisions. The problem associated with not reporting the planned sample size in a survey study is that, without this information, it is difficult for other researchers to evaluate the usefulness of the findings reported in the study. Even if the achieved sample size is good enough to justify the results of statistical analyses, there still remains a

BEST COPY AVAILABLE

question about how well the findings are generalizable. For instance, a sample size of 100 may be adequate for a hypothesis testing or an estimation of the population mean. There is a big difference, however, as to whether this sample size of 100 is obtained from a design in which 120 people were requested to respond, or obtained from a design in which 400 people were requested to respond . Such knowledge is definitely important to anyone who is interested in the results of your study. This also gets into response rate issue to be discussed later in this paper.

Another observation here is that one needs to be aware of what sampling scheme to use when determining the sample size in a survey study. The sample size tables that one typically finds in a reference text, such as that in McNamara's book (1994, p. 3), usually assume simple random sampling. In practice, true simple random samples are not often used, other sampling schemes, such as stratified or clustered sampling methods, are frequently employed instead. When these complex sampling methods, meaning methods other than the simple random sampling, are involved, it is not advised to use a sample size table mentioned above to select a sample size. Depending on the actual sampling method, different formulas are needed for sample size calculation.

Guideline Four: Selection Procedure Description

A clear description of the sample selection procedures implemented is an integrated component of sample design in a survey. In reporting quantitative research findings, it is a common practice to present detailed information about how the research, say an experiment, is conducted at every step. Such information serves two purposes. One is to let other researchers have sufficient information to evaluate the research and findings. The other is to provide adequate information so replication under comparable conditions by other researchers is possible (American Psychological Association, 1995).

In large-scale national surveys conducted by survey specialists, sample selection procedures are usually described in great detail in technical reports. In academic publications, journal space limit may not allow presenting every all the technical details. However, the space limit does not mean no disclosure at all about how a sample is actually selected. In fact, for a description of selection procedures, a few lines will be enough if a sample design is not very complex, which is usually the case in many evaluation survey studies.

Lack of an adequate description of sample selection procedures is frequently found in survey research publications. One may find such things as achieved sample size, demographics, and the type of sampling scheme in the methods section in a report. Little, however, is said about how the sample is selected from a population. In the 53 sample designs reviewed (Wang, 1996), about 10 appeared to have described their selection procedures with adequacy. The fact that 43 of the 53 sample designs did not provide adequate information about their selection procedures suggests a rather popular yet unjustified practice in reporting survey research findings. This practice is also found elsewhere (e.g., Pena & Henderson, 1986).

Without enough information about the selection procedures in obtaining a sample, it is hard to evaluate the usefulness of the sample in estimating a population characteristic. A major problem lies in the fact that there is no way to assess whether and how selection bias may affect the nature of the achieved sample. For example, selecting a true simple random sample assumes that a researcher has access to every member in the population and everyone is willing to cooperate. This seldom happens in reality. In a large scale survey that may involve several school districts or states, things become even more complicated and there has to be various compromises between the desired and achieved samples. When this occurs and no description is provided about the actual sample selection process, how can other researchers know what happened in between and to judge the value of the data and the findings?

Guideline Five: Response Rate and Non-response Treatment

A serious issue for survey research is the level of participation by sampled subjects. The participation level is measured by response rate. Although non-response and its potential effects seem to be the concern of estimation, data analysis and generalization, the problem should be considered at every phase of sample design, from population definition, sample size decisions, to selection, estimation and interpretation. Information about non-response rates and non-respondents' characteristics should also be disclosed, as is commonly emphasized in survey research literature (Aiken, 1988; Gall, Borg, & Gall, 1996; Jaeger, 1988; Shultz & Luloff, 1990; West, 1991).

The non-response problem has to be handled in both sample selection and estimation procedures of sample design. A researcher should always strive for high response rates when selecting a sample, which also means to minimize non-response rates. The non-response rate itself is not as simple as one might think. Although people tend to think of non-response rate in terms of unit non-response rate for the whole sample, non-response also occurs, and should be investigated, at the level of item non-response or subgroup non-response rates. For example, if everyone in a sample responded to a survey to yield a perfect response rate, then it was found that some items in the completed questionnaires were left blank for some reason (say, most of the items asked sensitive questions which few subjects wanted to give information about). This item non-response would pose a dilemma for the researchers in this case.

On the other hand, suppose a survey was designed to study compensation inequality between male and female administrators, and a stratified sample for both genders is selected. The data showed that the overall response rate was high, say, 80%. Further examination, however, found that one gender group had a near perfect return rate while the other had a low response rate. If not properly addressed, the difference in the subgroup non-response rates would also be a threat to the validity of

BEST COPY AVAILABLE

the findings. In a word, the non-response rate is a crucial piece of information with which the quality of data and validity of findings can be evaluated. This explains why such information needs to be made available to other researchers.

A related issue is the treatment of non-response if non-response rate is high enough to become a problem. A comparison between non-respondents and respondents on variables of interest is an important step to know one's sample and data quality. This information is also of interest to other researchers doing similar investigations, and should be disclosed if available. It is helpful to remember, however, a detailed sample description is not the same as a proper treatment of non-response. Sample description has more to do with population coverage in that such a description may show how comparable the sample and its population are on variables of interest. Investigation of non-respondent characteristics, on the other hand, addresses possible bias problems resulting from refusal to respond for certain reasons. It should be stressed that investigation of non-respondents' characteristics is always necessary (Aiken, 1988) and can not be spared. Generally speaking, non-response problem should be addressed in both selection process and in data analysis. During the stage of selection, aggressive follow-up efforts need to be made to encourage or persuade people to participate. In data analysis, adjustments can be made to reduce the influence of non-response.

The information from the 53 sample designs reviewed (Wang, 1996) indicates that most of the studies did mention the response rates achieved, but few discussed how non-response issue was handled. On the one hand, disclosing response rate is always a good practice as it gives readers an important piece of information about the quality of the sample and the study. On the other hand, it is not acceptable not to investigate non-response characteristics. In some of the studies evaluated, the response rate was rather low (50% or lower). Without any explanation about the nature of non-response and non-respondents, the survey findings are open to questions: some response bias is likely

to be present in the data from the achieved sample but the bias is not explored. A possible consequence from this situation is that the findings may not even apply to the originally-planned sample, let alone the population that the sample is supposed to represent.

Guideline Six: Estimation Procedures

Estimation procedures in a survey design may include such aspects as weighting, variance estimation, design effect, confidence interval estimation and treatment of non-response bias. That estimation is considered a criterion in sample design evaluation is because estimation procedures are closely related to the selection procedures and the data, and it is the outcome of the sample selection.

Weighting. This is an important strategy used in survey data analysis to adjust for unequal selection probabilities that result from complex sample designs; it is also used in post-stratification, and in making adjustment for total non-response (Kalton, 1983a, 1983b; Lee, Forthofer, & Lorimor, 1989). One typical instance of unequal selection probability occurs in stratified sampling where, intentional over-sampling of a particular stratum is desired to secure enough cases from the stratum for a reliable estimation. This happens when both total population and sub-population (stratum) parameter estimates are needed, but a sub-population may only be a small proportion of the total population.

Suppose we want to estimate teacher's job satisfaction in a school district of 1,000 teachers and we are also interested in the responses of the minority teachers in the district. We know that the minority teachers are about only six percent of the teacher population in the district. With an ideal simple random sample of 100 from the population, we will end up with about six minority teachers and 94 non-minority ones. Although 100 teachers may satisfy one's need to estimate the characteristics of the teacher's population as a whole, the group size of six is too small to be informative about the minority teachers as a sub-population. Therefore, we need to stratify the population by designating this particular minority group as a stratum and over-sample this group.

For example, we may sample 30 minority teachers from this stratum of 60, the selection probability is 0.5 (30 from 60), whereas the selection probability for the rest of the population is about .072 (70 from 970). When estimating the whole teacher population, each minority teacher may be assigned a weight of 2 (i.e., inverse of selection probability: 2=1/0.5) and each of the rest has a weight of 13.89 (inverse of selection probability: 13.89= 1/.072). This weighting essentially says that each minority teacher in the sample represents 2 minority teachers in the population, while each non-minority teacher in the sample represents 13.89 non-minority teachers in the population[1]. Separate analyses can then be done on the 30 minority teachers to explore the characteristics of this group.

Complex variance. When cluster sampling method is used in design and selection, variance estimation method assuming a simple random sample is not appropriate. Simply put, a cluster sample of size n usually yields a larger variance than does a simple random sample of the same size. This is mainly due to what Kish (1965) called the intraclass correlation among the responses within a cluster (e.g., students from a class). The effect of a cluster sampling design on variance is measured by "design effect", which is the ratio of the actual cluster sample variance estimate to the simple random sample variance estimate for a given same sample size n. If a cluster sample is obtained, but the analysis is performed using the method for simple random sample data, the result is usually an underestimation of the actual variance. In practice, the complex variance is usually estimated by approximation using a variety of resampling strategies (Lee et al., 1989).

Complex sample designs of one kind or another are frequently seen in literature involving survey. These designs include not only cluster sampling but also multi-stage sampling that may involve disproportionate selections for different sub-populations. However, few studies of such designs show

---

[1] For analysis purpose, one can assign a different set of weights (e.g., 0.20 and 1.39) to the two sample groups, as long as the ratio of the two weights remain the same.

any indications that such things as weighting or complex variance estimation were ever used or even considered. Typically, a research report describes sample, data collection method and analyses in such a way that it gives the impression that only a strictly simple random sample is used. The data analyses are usually done with some commercial statistical computing packages (e.g., SPSS, SAS) where computation is based on the assumption of using simple random sample data. When the data from a complex sample design is subjected to such a statistical analysis, the variance estimates are underestimated. Consequently, in hypothesis testing, the statistical power is falsely enhanced and a statistically significant result is more likely to occur at a given error rate. In parameter estimation, the underestimation of variance would lead to the construction of unduly narrow bounds around the point estimates, giving the false impression of great precision in estimation. All these are the problematic areas to look for when reading a survey research report.

It has been difficult for practitioners to handle the issue of design effect in cluster sampling design because of a lack of statistical software designed for the purpose. More recently, however, special procedures for handling survey data from complex cluster sampling designs have been incorporated in some statistical software. For example, the widely used SAS System, in its new Version 7 and Version 8 (Developer's Version), has incorporated several special procedures (e.g., SURVEYMEAN, SURVEYREG) for handling data from cluster sampling design (SAS Institute, 1999). Some special software designed specifically for this purpose have also been developed (e.g., SUDAAN, available the Research Triangle Institute in North Carolina). Currently, the software with the most extensive procedures for handing design effect of cluster sampling is probably STATA, which has incorporated more than a dozen special procedures ranging from estimating the standard error of sample mean to logistic regression analysis (StataCorp, 1999).

The easier availability of these special procedures to researchers will undoubtedly increase the awareness of the issue, and ultimately lead to better analytic strategies.

Non-response adjustment. Possible estimation bias due to non-response needs investigation. A survey population can be practically viewed as consisting two sub-populations, a respondent population and a non-respondent population. The decision to become a non-respondent or a respondent is seldom random. Therefore, there can be distinct differences between the two populations on the variables of interest. The purpose of survey is to obtain accurate information about the whole population, whose quantity can be expressed as the sum of the weighted values from the two populations:

$$\underline{Q} = \underline{w}_1\underline{q}_1 + \underline{w}_2\underline{q}_2$$

where $\underline{Q}$ is the population value of interest, $\underline{q}_1$ is the known sample response value, $\underline{q}_2$ is the unknown non-response value, and $\underline{w}_1$, $\underline{w}_2$ are the weights for the percentage of population of the respondent and non-respondent sub-populations respectively. The estimation of $\underline{Q}$ depends on both $\underline{w}_2$ and $\underline{q}_2$. Without a direct investigation, $\underline{q}_2$ is always unknown. The best way to ensure a reasonable $\underline{Q}$ is to keep $\underline{w}_2$ as small as possible to minimize the impact of the non-respondent population on the whole population.

Assessment of non-response effects and possible adjustments are to find out what $\underline{q}_2$ would be if the non-respondents had responded. Because no one ever knows what $\underline{q}_2$ is, we have to rely on some reasonable approximation. A host of techniques are available in sample survey literature. Typically, respondents and non-respondents are compared on some ancillary information that is thought to be related to survey purposes and may influence one's response to survey questions. Various weighting and imputation methods are also proposed to compensate for both unit non-response and item non-response.

Summary and Conclusions

When it comes to evaluating the quality of survey study findings, one needs to examine the sample design part with great care because the design determines the sample selection and estimation procedures. Although there are many things to look for in a sample design, the six criteria discussed in this paper are probably the most important for consumers of survey research to keep in mind when reading a report of survey research.

The six key criteria covered both selection and estimation in general. In particular, a good survey sample design should have information about a clearly specified population of interest. This information relates to the validity and generalizability of the findings of the study.

The unit of analysis should be explicitly stated in the design so that the relevance of the chosen unit of analysis to the research questions of the survey study can be checked. It is always helpful to ask whether the correct unit of analysis is used in analysis for the given purposes of a study.

It is not enough to just let readers know the actual sample size used in a study. The planned sample size and how the size is determined need to be explained in a report. Without this information, readers are unable to know why a certain sample size is chosen over other possibilities or whether such aspects as statistical power, margin of error, design effect, etc., have ever been taken into account.

The procedures with which a sample is actually selected help readers not only evaluate the quality of the sample, but also plan for similar sample selection in their own survey studies. So adequate description of the selection procedures is important.

Non-response causes headache to survey researchers and has to be dealt with with great efforts. Treatment of non-response can be done with good follow-up efforts to minimize non-response rate in the selection stage. In data analysis stage, necessary adjustments can be made to reduce the

impact of the non-response. Such information should be made available to readers so that they can better understand the meanings of the survey results as reported.

Finally, several things call for consideration in estimation. These are weighting, complex variance estimation, and non-response adjustment. These procedures are often needed because true simple random samples are seldom used, and non-response of different degrees is almost a reality in most survey studies. Without well-planned estimation procedures, analysis results can be affected. It is very important to know how the statistics or population estimates are obtained from data and, given this knowledge, how justifiable and useful the findings may be.

Reference

Aiken, L. R. (1988). The problem of non-response in survey research. Journal of Experimental Education, 56(3), 116-119.

American Psychological Association. (1995). Publication manual of the American Psychological Association (4th ed.). Washington, DC: Author.

Babbie, E. (1986). The practice of social research (4th ed.). Belmont, CA: Wadsworth.

Babbie, E. (1990). Survey research methods (2nd ed.). Belmont, CA: Wadsworth.

Barnett, V. (1991). Sample survey principles and methods. New York: Oxford University Press.

Bryk, A. S., & Raudenbush, S. W. (1992). Hierarchical linear models: Application and data analysis methods. Newbury Park, CA: Sage.

Cochran, W. G. (1977). Sampling techniques. New York: John Wiley & Sons.

Cohen, J. (1988). Statistical power for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Fink, A. (1995). How to sample in surveys. Thousand Oaks, CA: Sage.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). Educational research (6th ed.). New York: Longman.

Jaeger, R. M. (1988). Survey research methods in education. In R. M. Jaeger (Ed.), Complementary methods for research in education (pp. 303-330). Washington, DC: American Educational Research Association.

Jolliffe, F. R. (1986). Survey design and analysis. New York: John Wiley & Sons.

Kalton, G. (1983a). Compensating for missing survey data. Ann Arbor, MI: Institute for Social Research, University of Michigan.

Kalton, G. (1983b). Introduction to survey sampling. Beverly Hills, CA: Sage.

Kish, L. (1965). Survey sampling. New York: John Wiley & Sons.

Lee, E. S., Forthofer, R. N., & Lorimor, R. J. (1989). Analyzing complex survey data. Newbury Park, CA: Sage.

McNamara, J. F. (1994). Surveys and experiments in education research. Lancaster, PA: Technomic.

Miskel, C. & Sandlin, T. (1981). Survey research in educational administration. Educational Administration Quarterly, 17(4), 1-20.

Pena, D. M. & Henderson, R. D. (1986). Sampling procedures used for national surveys of public school teachers -- problems and possible solutions. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Permut, J. E., Michel, A. J., & Joseph, M. (1976). The researcher's sample: A review of the choice of respondents in marketing research. Journal of Marketing Research, 13, 278-283.

SAS Institute. (1999). What's new in SAS software for Version 7 and the Version 8 developer's release? Cary, NC: SAS Institute, Inc.

Scheaffer, R. L., Mendenhall, W., & Ott, L. (1990). Elementary survey sampling (4th ed.). Boston, MA: PWS-Kent.

Shultz, S. D., & Luloff, A. E. (1990). The threat of non-response bias to survey research. Journal of the Community Development Society, 21(2), 104-115.

StatCorp. (1999). Stata Reference Manual Release 6, Volume 4 Su-Z. College Station, TX: Stata Press.

Sudman, S. (1976). Applied sampling. New York: Academic Press.

Thompson, S. K. (1992). Sampling. New York: Wiley.

Wang, L. (1996). A typology and evaluation of the survey sample designs in the Educational Administration Quarterly: 1980-1995. Unpublished doctoral dissertation, Texas A&M University, College Station.

Wang, L., & McNamara, F. J. (1997). An Evaluation of the Sample Designs in Educational Survey Research. Paper presented at the 1997 annual meeting of the American Educational Research Association, Chicago, IL, March, 1997.

West, L. J. (1991). A frequently violated requirement in survey research. Journal of Education for Business, 66(3), 134-135.